INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Improving Joint Speech and Emotion Recognition Using Global Style Tokens

Jehyun Kyung, Ju-Seok Seong*, Jeong-Hwan Choi, Ye-Rin Jeoung, and Joon-Hyuk Chang*

Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea
{jehyunkyung, as2835510, brent1104, jyr0328, jchang}@hanyang.ac.kr

Abstract

Automatic speech recognition (ASR) and speech emotion recognition (SER) are closely related in that the acoustic features of speech, such as pitch, tone, and intensity, can vary according to the speaker's emotional state. Our study focuses on a joint ASR and SER task, in which an emotion token is tagged and recognized along with the text. To further improve the joint recognition performance, we propose a novel training method that adopts the global style tokens (GSTs). The style embedding is extracted from the GSTs module to enhance the joint ASR and SER model to capture emotional information from speech. Specifically, a conformer-based joint ASR and SER model pre-trained on a large-scale dataset is jointly fine-tuned with style embedding to improve both ASR and SER. The experimental results on the IEMOCAP dataset showed that the proposed model achieves a word error rate of 15.8% and four emotion classification weighted and unweighted accuracy of 75.1% and 76.3%, respectively.

Index Terms: automatic speech recognition, speech emotion recognition, global style tokens

1. Introduction

Automatic speech recognition (ASR) and speech emotion recognition (SER) are closely related research areas that have gained significant attention in recent years [1, 2]. ASR focuses on transcribing speech into text, while SER aims to recognize a speaker's emotional state based on speech signals. ASR has made considerable progress in recent years, thanks to the availability of large amounts of data, advances in training techniques and efficient structures. Especially, conformer-based models [3, 4] which show impressive results of speech recognition benchmark datasets by capturing both local and global features. Despite of these advance, ASR still faces challenges recognizing speech from speakers with different accents, languages, and emotional states [5, 6].

Researchers have explored various approaches in the area of SER [7, 8], including the use of acoustic features such as pitch, loudness, and spectral information to extract emotional information from speech signals. Some studies [9, 10] have also employed the linguistic features, obtained through sentiment analysis, to improve the accuracy of SER models. However, despite significant progress, SER also still faces challenges, such as dealing with individual differences in emotional expression and identifying emotions in noisy environments. Several studies [11–13] have explored the integration of ASR and SER to improve the performance of both tasks. One approach [14] is to use the latent features extracted from a pre-trained ASR model to perform SER. Concurrently, some studies [15, 16] have proposed using an emotion-dependent feature selection method to select the most informative acoustic features for ASR based on the speaker's emotional state. Other approaches [6, 17, 18] use text outputs from the pre-trained ASR models, leading to improved SER accuracy.

Our proposed method follows the recent trends in joint ASR and SER task, where a unified ASR–SER model is trained to simultaneously recognize the text and emotional state of the speech signals. Specifically, we aim to improve the performance of both ASR and SER by incorporating global style tokens (GSTs) [19] during the training. The GSTs module adds an additional input to the desired system, which is a set of learned embedding called "style embedding". This style embedding captures the global characteristics of the speaking style, such as emotion, pitch, speaking rate, or intonation. By conditioning the style embedding into our proposed model, we aim to reduce the word error rate (WER) of emotional speech and improve the accuracy of the speaker's emotional state prediction. The experimental results demonstrated the effectiveness of our approach in enhancing the WER of emotional speech and recognizing the speaker's emotional state. The main contributions of our work are summarized as follows:

- Based on a pre-trained ASR model using a large-scale speech dataset, we propose a joint training method for ASR and SER in an emotional speech.
- We propose a novel training method that adopts the GSTs to improve the joint ASR and SER performance.
- Our proposed approach achieved the state-of-the-art ASR and SER performance on the IEMOCAP dataset.

2. Related Works and Motivations

Kons *et al.* [5] developed a recurrent neural network (RNN)-transducer-based ASR model that can perform both ASR and SER simultaneously. To enable joint ASR and SER, an emotion token was added to the text transcription. However, the joint model performed worse than separately trained ASR and SER models. Kons *et al.* explained that this was because the original training of the ASR model ignored a lot of information in the speech useful for accurately identifying emotions. Therefore, further research is needed, either by using a more extensive training dataset or incorporating the emotion classification objective into the training of the ASR. Chen *et al.* [6] used a conformer-based encoder pre-trained with a large ASR dataset for the SER task. The embedding extracted from the encoder was introduced to a multi-head self-attention (MHSA)-based RNN structure to perform emotion recognition. Chen *et al.* showed that a model pre-trained on the large-scale corpus dataset used in the ASR task could leverage the SER task.

The GST-Tacotron model is a neural network-based ap-

*Equal contributions.

10.21437/Interspeech.2023-2375

proach to generate speech that can sound more natural and expressive [20]. This model combines the Tacotron architecture, which is an end-to-end speech synthesis system [21], with a novel technique called GSTs. The role of the GSTs is to help the system capture and model the global speaking style characteristics that are not directly related to the input. Inspired by the two previous researches [5, 6], we aim to develop an efficient joint ASR and SER model for emotional input speech. Specifically, we use a conformer-based attention-based encoder-decoder (AED) model [22] pre-trained on a large speech recognition dataset as the baseline model for our study. By incorporating the GSTs into the baseline model, our proposed model aims to improve emotional speech transcription and emotion recognition performance.

3. Method

3.1. Joint speech and emotion recognition

The overall framework we propose for the joint ASR and SER model is illustrated in Figure 1. We use the conformer-based AED model, depicted on the left side of the Figure 1, initialized by applying a pre-trained model to obtain efficient convergence and optimal performance for joint ASR and SER training. The AED model have been used in typical end-to-end ASR, which outperforms the connectionist temporal classification (CTC) model [23]. Because previous studies have suggested that multi-task learning using the CTC loss function can improve the performance of the AED model, we design the overall loss function as a weighted sum of AED and CTC loss functions:

$$\mathcal{L}_{Overall} = -\frac{1}{N}\sum_{n=1}^{N}\log\big(\alpha\mathcal{P}_{AED}(\mathbf{c}_n|\mathbf{x}_n) + (1-\alpha)\mathcal{P}_{CTC}(\mathbf{c}_n|\mathbf{x}_n)\big), \quad (1)$$

where α is a hyperparameter that satisfying $0 \leq \alpha \leq 1$, and $\mathbf{x} \in \mathbb{R}^{T'}$ and $\mathbf{c} \in \mathbb{R}^{L+1}$ denote input speech utterances and target sequence of its labels, respectively. We denote the corresponding word sequence of $\mathbf{x}$ as $\mathbf{c}' \in \mathbb{R}^{L}$, where T', L, and N is a length of raw speech, word sequence, and batch size, respectively. Our target word sequence is defined by additionally including a corresponding emotion token e of $\mathbf{x}$ as: $\mathbf{c} = \mathbf{c}' \cup \{e\}$, where $e \in \{e_1, ..., e_n\}$ of n-emotional classes. Finally, $\mathcal{P}_{AED}$ and $\mathcal{P}_{CTC}$ denote the posterior probability of the $\mathbf{c}$ conditioned on the input, $\mathbf{x}$ for AED and CTC loss, respectively. Specifically, $\mathcal{P}_{AED}$ can be formulated by using the additional state token, end-of-sentence symbol, $\langle\text{eos}\rangle$, as follows:

$$\mathcal{P}_{AED}(\mathbf{c}|\mathbf{x}) = \prod_{i=1}^{L+2}\mathcal{P}\big(c_i|c_{i-1}, \ldots, c_0, \mathbf{v}_i\big), \quad (2)$$

where $c_{L+2} = \langle\text{eos}\rangle$, $c_{L+1} = e$, $c_0 = \langle\text{sos}\rangle$, and $\mathbf{v}$ denotes as a context vector, which aggregates the relevant portions of the encoder and decoder outputs within attention method. The start-of-sentence symbol, $\langle\text{sos}\rangle$, which serves before the decoder produces any outputs, it is used as decoders initial input. $\mathcal{P}_{CTC}$ denotes the posterior probability for CTC, by marginalizing over all possible alignments for CTC, and it can be formulated as:

$$\mathcal{P}_{CTC}(\mathbf{c}|\mathbf{x}) = \sum_{\mathcal{A}\in\mathcal{A}_{\mathbf{c},\mathbf{x}}^{CTC}}\prod_{t=1}^{T}\mathcal{P}\big(a_t|a_{t-1}, \ldots, a_1, \mathbf{h}^M = (h_t^M, \ldots, h_1^M)\big), \quad (3)$$

where $a_T = e$, and $\mathcal{A}_{\mathbf{c},\mathbf{x}}^{CTC}$ denote the set of all valid alignments, each alignment of time frame t is denoted as a_t. Also,

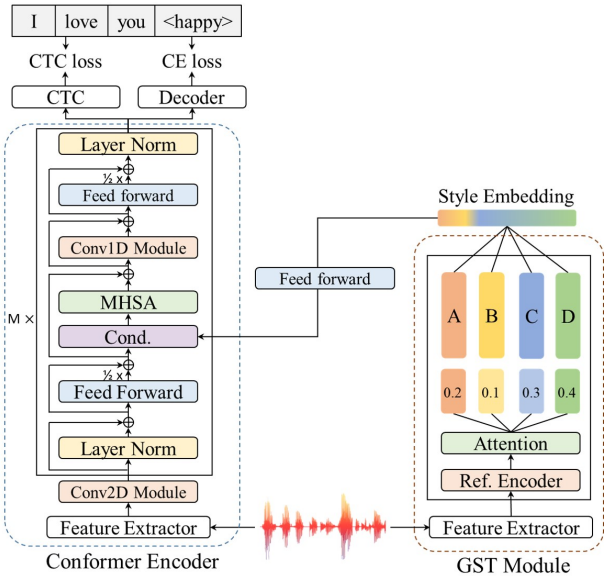


Figure 1: *Schematic of proposed joint model with GSTs module.*

$\mathbf{h}^m \in \mathbb{R}^{T\times D}$ denotes the output of the m-th conformer encoder block, and the M, T, D are a number of encoder blocks, frame length, and representation feature size of $\mathbf{h}^m$, respectively. Our baseline model, based on the conformer-based AED model initialized by a pre-trained model and fine-tuned by Equation 1, is an extended method compared to the previous joint ASR and SER research. Especially, using a pre-trained model improves the training process through efficient convergence, and placing an emotion token at the end of the word sequence could utilize the decoder's ability of AED, resulting in classifying emotion states more accurately.

3.2. Conditioning style embedding

We use the GSTs module, which is commonly used in speech synthesis tasks. Because of its emotional modeling property, we use it when jointly training emotional speech and the speaker's emotional state. The style embedding, $\mathbf{s}_e \in \mathbb{R}^{D'}$, is the output of the GSTs module $\mathbf{s}_e = \text{GSTs}(\mathbf{x})$ for the speech $\mathbf{x}$. After that, we use one feed-forward network (FFN) to add it to the outputs of one half-step FFN of every conformer encoder layers. The detailed formulas mentioned above are as follows:

$$\tilde{\mathbf{h}}^m = \mathbf{h}^m + \frac{1}{2}\text{FFN}\left(\text{LayerNorm}\left(\mathbf{h}^{m-1}\right)\right), \quad (4)$$

$$\hat{\mathbf{h}}^m = \tilde{\mathbf{h}}^m + \text{FFN}(\mathbf{s}_e), \quad (5)$$

where $\tilde{\mathbf{h}}^m$ and $\hat{\mathbf{h}}^m$ denote hidden representations of m-th the conformer encoder block and those dimension are in $\mathbb{R}^{T\times D}$, respectively. As illustrated in Figure 1, the last parts of the conformer encoder block is organized same with original conformer as in [3]. The style embedding, which is mapped by one FFN module, is used to condition multiple times to all the encoder blocks.

4. Experiments Setup

4.1. Dataset and preparation

We employed a large-scale pre-trained AED model[1] as our baseline, which was trained on a 10,000 hours subset of Gi-

[1]Pre-trained conformer-based AED model can be downloaded from https://zenodo.org/record/4630406#.ZAgdhnZByUl.

Table 1: *WER of the proposed method on IEMOCAP dataset.*

Method	WER (%)				WW (%)	UW (%)
	Hap	Sad	Ang	Neu		
Conformer-based AED	20.1	15.5	12.0	19.0	17.3	16.6
+ emotion labels cond.†	17.6	13.8	10.3	17.2	15.3	14.7
+ GSTs cond.	19.9	14.7	11.6	18.9	17.0	16.3
+ P.T. GSTs cond.	19.1	14.6	11.3	18.6	16.5	15.9

†Indicates the oracle one-hot emotion labels.

gaspeech dataset [24]. Gigaspeech dataset is a multi-domain English ASR corpus that includes recordings from various environments, such as audiobooks, podcasts, and YouTube.

We used the IEMOCAP dataset [25] to train and evaluate the ASR and SER for the proposed method. The IEMOCAP dataset is a multi-modal dataset containing 12 hours of audio, text, emotion label, and video recordings of naturalistic dyadic conversations of 10 actors (5 male, 5 female) in English. The dataset consists of 10 sessions, each session containing a conversation between two actors. We merged excitement and happiness into the single happiness emotion class to make fair comparison with the previous studies [26]; the four emotions (n=4) were used (happy, sad, angry, and neutral). There were a total of 5,531 utterances across all sessions. The label distribution of the dataset was imbalanced, with neutral being the most frequent (30.9%), followed by happiness (29.6%), anger (19.9%), and sadness (19.6%).

4.2. Training procedure and evaluation

We applied 5-fold cross-validation for the IEMOCAP dataset to prevent the impact of limited data. We used three sessions for training, one for validation, and the remaining for testing. We used the Adam optimizer [27] with a learning rate of $1.5e$-3 to train the proposed joint ASR and SER model. We trained the model for 100 max epochs with the early stopping strategy depending on the evaluation performed on the validation set. We used batch size of 64 (N=64) and set the hyperparameter α = 0.3. For evaluation, we chose the 10-best models according to the validation accuracy. We averaged the 10-best models when decoding the validation and the test sets. We used the beam search with a beam size of ten. We implemented the proposed method in PyTorch [28] and conducted experiments on two NVIDIA GeForce RTX 3090 GPUs with 24GB memory.

4.3. Evaluation metrics

We evaluated the WER and emotion prediction accuracy for joint ASR and SER tasks. Because the IEMOCAP dataset suffers from class imbalance, the emotion prediction accuracy was evaluated using weighted accuracy **(WA)**: accuracy over all classes; and unweighted accuracy **(UA)**: the average accuracy for each category [29]. For the same reason, we evaluated ASR utilizing weight with WER. We denoted the weighted WER **(WW)**: WER over all classes; and unweighted WER **(UW)**: the average WER for each category. To the best of our knowledge, **WW** and **UW** are the first attempts. With the exception of Tables 3 and 5, which will be discussed later, all experiments in this study used text transcription with emotion token to calculate WER.

4.4. Model architecture

We used the ESPnet toolkit[2] [30] to implement the AED model. The AED model consisted of 12 layers (M=12) of the conformer-based encoder and six layers of the transformer-based decoder [3, 4] identical to those described in Chen *et al.* [6]. The number of parameters in the AED model was 112M. The recognition unit was composed of words, and each emotion token was treated as a unique recognition unit to be output after the speech recognition result was produced. We used the pre-trained GSTs in the GST-Tacotron model[3] trained on the emotional dataset for the text-to-speech task. The GSTs module consisted of a reference encoder and a style token layer. The reference encoder consisted of six layers of convolution and one layer of gated recurrent unit network. The style token layer received the output of the reference encoder as an input and outputs the 256-dimensional style embedding through the attention module. The number of parameters in the GSTs module was about 480k. The feature extractors for the AED model and GSTs module converted raw input speech into an 80-dimensional log-mel spectrogram using a Hann window. However, there were several differences in their configurations. The AED model used a 512-size FFT, 32 ms window length, and 16 ms hop size, while the GSTs module used a 2048-size FFT, 50 ms window length, and 12.5 ms hop size.

[2]ESPnet toolkit source code from https://github.com/espnet/espnet.

Table 2: *Accuracy of emotion prediction of the proposed method for IEMOCAP dataset.*

Method	ACC (%)				WA (%)	UA (%)
	Hap	Sad	Ang	Neu		
Conformer-based AED	67.7	77.8	78.8	71.1	72.8	73.9
+ emotion labels cond.†	100	99.9	100	100	100	100
+ GSTs cond.	65.4	78.3	82.7	72.1	73.2	74.6
+ P.T. GSTs cond.	70.1	79.7	82.9	72.4	75.1	76.3

†Indicates the oracle one-hot emotion labels.

Table 3: *Comparison of WER with and without emotion token in the output of the proposed model.* ***W/ E.T.*** *indicates the with emotion token.*

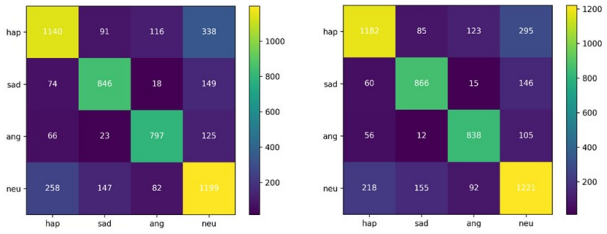
Method	W/ E.T.	WER (%)				WW (%)	UW (%)
		Hap	Sad	Ang	Neu		
Conformer-based AED	✓	20.1	15.5	12.0	19.0	17.3	16.6
+ P.T. GSTs cond.	✓	19.1	14.6	11.3	18.6	16.5	15.9
Conformer-based AED	✗	19.1	14.9	11.3	18.1	16.4	15.8
+ P.T. GSTs cond.	✗	18.3	14.1	10.8	17.7	15.8	15.2

5. Experimental Results

5.1. Effect of style embedding conditioning

We first evaluated joint ASR and SER performance on the conformer-based AED model without conditioning on style embedding. Then, we validated our experiments by conditioning on the conformer-based AED model with the one-hot encoding of the target emotion labels, which denote an oracle environment. Tables 1 and 2 show the WER and emotion prediction accuracy, including performances for each of the four emotion classes, respectively. As shown in the first and second lines of each table, all performance for the oracle environment was better than the baseline. Thus, we compared the performance changes for ASR and SER when conditioning the style embedding extracted from non-pre-trained and pre-trained GSTs module to the conformer-based AED model, respectively. The method of conditioning the style embedding extracted from the non-pre-trained GSTs module with the conformer-based AED model improved the performance, but it was not significant. The method of conditioning the style embedding extracted from

[3]Pre-trained GST-Tacotron model and the source code were from https://github.com/KinglittleQ/GST-Tacotron.



(a) Conformer-based AED (b) Baseline w/ P.T. GSTs cond.

Figure 2: *Confusion matrix for the emotion recognition performance. The x-axis and y-axis represent the prediction and reference emotion labels.*

the pre-trained GSTs module with the conformer-based AED model showed a 4.6% and 4.2% performance improvement in **WW** and **UW**, respectively, compared to baseline performance. Similarly, **WA** and **UA** were improved by 3.2% and 3.2% relatively. We consider style embedding extracted from pre-trained GSTs modules useful for emotion prediction and improving transcription of emotional speech.

We also investigated the WER calculated by excluding emotion token from the output of our proposed model trained with text including emotion token and reported it in Table 3. Our proposed model showed 15.8% when calculating the WER without emotion token on the IEMOCAP dataset. This result is noteworthy when compared to other studies. Figure 2 shows the confusion matrix for emotion prediction performance in the baseline and proposed experimental environments. The proposed simultaneous ASR and SER model showed improved prediction performance for all four emotion classes, particularly for the happy emotion class.

5.2. The analysis of conditioning encoder layers

We examined the effect of conditioning different layers of the encoder in a speech and emotion recognition system. Specifically, we analyzed the trends in WER as we conditioned each encoder layer. Figure 3 illustrates that the WER gradually improved as we conditioned on the lower encoder layer compared to the upper ones. Notably, the best performance was achieved when we conditioned all the layers. Our findings suggest that conditioning on multiple layers of representation can effectively improve the accuracy of speech recognition.

Table 4: *Comparison of performance based on emotion token position.* ***L.E.*** *indicates whether the emotion token is located front or end of the text transcription.*

Method	L.E.	SER		ASR	
		WA (%)	UA (%)	WW (%)	UW (%)
Conformer-based AED	front	71.3	72.6	18.3	17.8
	end	72.8	73.9	17.3	16.6
+ P.T. GSTs cond.	front	73.0	74.2	17.7	17.3
	end	75.1	76.3	16.5	15.9

5.3. Effect of emotion token position

We compared the performance variation of ASR and SER according to the position of the emotion token in Table 4. The experimental results showed that both ASR and SER perform better when the emotion token was positioned after than before the text transcription. Placing an emotion token in front of a text transcription could lead to incorrect transcription if the predicted emotion was inaccurate. However, when the emotion token was placed at the end of the text transcription, emotion recognition performance improved as contextual information was used to predict the emotion.

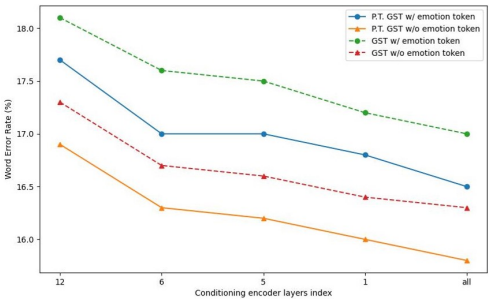


Figure 3: *Comparison of WER performance changes when conditioning style embeddings at specific encoder block indexes.*

5.4. Comparison to previous works

Table 5 shows our comparison of ASR and SER with previous works on the IEMOCAP dataset. Our method is extended from the two previous methods [5, 6], and we propose a method for conditioning the style embedding extracted from the GSTs module to the AED model. To the best of our knowledge, the proposed method achieves the state-of-the-art in both ASR and SER performance on the IEMOCAP dataset.

Table 5: *Comparing ASR and SER performance with previous works on the IEMOCAP dataset.* ***P.T.*** *indicates whether a large-scale pre-trained model is employed, and* ***J.T.*** *indicates whether ASR and SER are performed jointly. Our WER is calculated without emotion token in this table.*

Method	P.T.	J.T.	SER		ASR
			WA (%)	UA (%)	WER (%)
Previous					
Yeh et al. (2020) [12]			63.1	64.4	56.4
Feng et al. (2020) [11]			68.6	69.7	35.7
Amiriparian et al. (2021) [31]			-	73.8	22.2
Li et al. (2022) [13]			63.4	-	32.7
Kons et al. (2022) [5]		✓	72.0	-	20.8
Chen et al. (2022) [6]	✓		-	76.1	16.4
Proposed					
Conformer-based AED	✓	✓	72.8	73.9	17.3
+ P.T. GSTs cond.	✓	✓	**75.1**	**76.3**	**15.8**

6. Conclusions

Our study proposed the novel joint ASR and SER model to improve emotional speech recognition by conditioning the style embedding extracted from the GSTs module to the encoder layers. Our method allowed for a combined output of a speech signal's text transcription and emotional state, enabling efficient joint speech and emotion recognition. The experimental results on the IEMOCAP dataset showed that the joint model achieved a significantly improved WER and accuracy of emotion prediction in an emotional speech and demonstrated the effectiveness of incorporating style embedding conditioning into the joint model. Our proposed model showed significant potential to improve the performance of recognizing emotional speech.

7. Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-00456, Development of Ultra-high Speech Quality Technology for Remote Multi-speaker Conference System)

8. References

[1] C. Zhang and L. Xue, "Autoencoder with emotion embedding for speech emotion recognition," *IEEE Access*, vol. 9, pp. 51 231–51 241, 2021.

[2] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning," in *Proc. INTERSPEECH*, 2021, pp. 4508–4512.

[3] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH*, 2020, pp. 5036–5040.

[4] P. Guo *et al.*, "Recent developments on espnet toolkit boosted by conformer," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 5874–5878.

[5] Z. Kons *et al.*, "Extending RNN-T-based speech recognition systems with emotion and language classification," in *Proc. INTERSPEECH*, 2022, pp. 546–549.

[6] C. Chen and P. Zhang, "CTA-RNN: Channel and temporal-wise attention rnn leveraging pre-trained asr embeddings for speech emotion recognition," in *Proc. INTERSPEECH*, 2022, pp. 4730–4734.

[7] A. Tursunov, S. Kwon, and H.-S. Pang, "Discriminating emotions in the valence dimension from speech using timbre features," *Appl. Sci.*, vol. 9, no. 12, p. 2470, 2019.

[8] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," *Int. J. Speech Technol.*, vol. 23, no. 1, pp. 45–55, 2020.

[9] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, and A. Gelbukh, "Urdu sentiment analysis with deep learning methods," *IEEE Access*, vol. 9, pp. 97 803–97 812, 2021.

[10] Z. Nasim, Q. Rajput, and S. Haider, "Sentiment analysis of student feedback using machine learning and lexicon based approaches," in *Proc. Int. Conf. Res. Innov. Inf. Syst.*, 2017, pp. 1–6.

[11] H. Feng, S. Ueno, and T. Kawahara, "End-to-end speech emotion recognition combined with acoustic-to-word asr model," in *Proc. INTERSPEECH*, 2020, pp. 501–505.

[12] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "Speech representation learning for emotion recognition using end-to-end asr with factorized adaptation," in *Proc. INTERSPEECH*, 2020, pp. 536–540.

[13] Y. Li, P. Bell, and C. Lai, "Fusing asr outputs in joint training for speech emotion recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7362–7366.

[14] J. Santoso, T. Yamada, S. Makino, K. Ishizuka, and T. Hiramura, "Speech emotion recognition based on attention weight correction using word-level confidence measure," in *Proc. INTERSPEECH*, 2021, pp. 1947–1951.

[15] M. Abdelwahab and C. Busso, "Ensemble feature selection for domain adaptation in speech emotion recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 5000–5004.

[16] V. Hozjan and Z. Kačič, "A rule-based emotion-dependent feature extraction method for emotion analysis from speech," *J. Acoust. Soc. Am.*, vol. 119, no. 5, pp. 3109–3120, 2006.

[17] S. Sahu, V. Mitra, N. Seneviratne, and C. Y. Espy-Wilson, "Multi-modal learning for speech emotion recognition: An analysis and comparison of asr outputs with ground truth transcription," in *Proc. INTERSPEECH*, 2019, pp. 3302–3306.

[18] A. Ghriss, B. Yang, V. Rozgic, E. Shriberg, and C. Wang, "Sentiment-aware automatic speech recognition pre-training for enhanced speech emotion recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7347–7351.

[19] Y. Wang *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5180–5189.

[20] R. Skerry-Ryan *et al.*, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. Int. Conf. Mach. Learn*, 2018, pp. 4693–4702.

[21] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, 2017, pp. 4006–4010.

[22] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 577–585.

[23] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[24] G. Chen *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," in *Proc. INTERSPEECH*, 2021.

[25] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, pp. 335–359, 2008.

[26] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6484–6488.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[28] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.

[29] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7179–7183.

[30] S. Watanabe *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, 2018, pp. 2207–2211.

[31] S. Amiriparian *et al.*, "On the impact of word error rate on acoustic-linguistic speech emotion recognition: An update for the deep learning era," *arXiv:2104.10121*, 2021.